Student Evaluation

Alberta
EDUCATION

Report of
AD HOC STUDENT EVALUATION
TECHNICAL ADVISORY COMMITTEE

SECOND MEETING

FEBRUARY 21, 1983 -- 9:00 a.m.

4TH FLOOR, HARLEY COURT
EDMONTON, ALBERTA

Alberta
EDUCATION

Devonian Building, West Tower, 11160 Jasper Avenue, Edmonton, Alberta, Canada T5K 0L2

May 18, 1983

TO: ALL MEMBERS OF THE AD HOC STUDENT
EVALUATION TECHNICAL ADVISORY COMMITTEE

Re: Meeting of the Ad Hoc Student Evaluation
Technical Advisory Committee, Thursday, May 26, 1983

A meeting of the Ad Hoc Student Evaluation Technical Advisory Committee will be held on Thursday, May 26, 1983, in the Boardroom of the Student Evaluation Branch, 4th Floor, Harley Court, 10045 - 111 Street, commencing at 9:00 a.m.

The purpose of the meeting, as mandated, is to discuss the recommendations of the Committee with the Director of the Student Evaluation Branch prior to finalizing the report.

A copy of the revised report is enclosed.

Should you have any questions, please call me at 427-9399.

Sincerely,

Patricia Ford

Patricia Ford
Research Officer
Student Evaluation Branch

PF/bn

Tech. Ad.

Alberta
EDUCATION

Devonian Building, West Tower, 11160 Jasper Avenue, Edmonton, Alberta, Canada T5K 0L2

March 2, 1983

To: All Members of the Ad Hoc Student Evaluation
Technical Advisory Committee:

Re: Meeting of the Ad Hoc Student Evaluation
Technical Advisory Committee, March 14, 1983

The third meeting of the Ad Hoc Student Evaluation Technical Advisory Committee will be held on Monday, March 14, 1983 in the Boardroom of the Student Evaluation Branch, 4th Floor, Harley Court, 10045 - 111 Street. The meeting will commence at 9:00 a.m.

The proposed agenda includes the following items:

1. Sampling Techniques
2. Matrix Sampling
3. Standards and Setting of Standards
4. Test Security (Test Equivalency)

Please retain all receipts for expenditures related to your membership on this Committee (air fare, taxi charges, mileage, etc.), as Expense Claim Forms are prepared following every meeting.

If you have any questions, please call Patricia Ford at 427-9398.

Yours sincerely,

Patricia Ford

Patricia Ford
Research Officer
Student Evaluation Branch

PF/bn

Minutes of the Second Meeting
of the
AD HOC STUDENT EVALUATION TECHNICAL ADVISORY COMMITTEE
in the Boardroom of
Harley Court, 4th Floor
10045 - 111 Street, Edmonton, Alberta
February 21, 1983

Members Present

J. Lawrence Tymko (Chairman) (SEB)
William Brooks (ATA)
Janelle Holmes (SEB)
Michael Kozlow (SEB)
Thomas Maguire (U.of A.)
Harry Mosychuk (Edmonton Public School Board)
David Wasserman (SEB)

Absent

Larry Ferguson (ASTA)
Eric Mokosch (U. of L.)

Participants

Lloyd E. Symyrozum (Director, SEB)
Robert Runte (Test Development Specialist, Social Studies)

Recorder

Patricia Ford

AD HOC STUDENT EVALUATION TECHNICAL ADVISORY COMMITTEE

MINUTES OF FEBRUARY 21, 1983 MEETING

The meeting was called to order at 9:00 a.m.

1. THE REVISED STUDENT EVALUATION POLICY DISCUSSION PAPER

L. Symyrozum joined the meeting and asked external members on the committee for feedback regarding the discussion paper.

H. Mosychuk: questioned the "purpose" section of the paper, noting that the testing program should encompass all aspects of testing and should therefore address Appeal Examinations. There is no provision for Appeal Examinations in the "purpose" section.

He requested clarification re the Provincial Róle, pointing out that comprehensives essentially certify that students have acquired certain knowledge and skills as reflected in teacher-assigned marks. To reflect this, he recommended that the statement

> 1. *Certifying the basic skills and knowledge attained by students graduating from high school through the awarding of a High School Diploma.*

be changed to read

> 1. *Certifying the basic skills and knowledge attained by students graduating from high school through compulsory testing.*

J. Holmes: does not dispute the underlying premises, but feels the policy is determining ends. If ends are not met, do we go back to changing policy. Also feels there is no provision for individual needs at the local or jurisdictional levels.

H. Mosychuk: recommended that policy specify what portion of the student population is affected and what portion of that population is expected to pass and qualify for the High School Diploma. *We should be setting standards.*

W. Unruh: committee has to clarify whether this is a competency-based approach. It is important that standards be set and it is important to clarify the purpose of the marks. He emphasized the need to ensure that methods and design fit the goals of the particular test so as not to confound methodology and context.

The committee agreed that standards should be set in place in this regard.

The committee adjourned for coffee from 10:00 to 10:15 a.m.

.../2

2. TEST DEVELOPMENT PROCESS

Several documents which had been circulated were presented as information items:

- Item Control Sheet
- Test Blueprint
- Proposed Categories for Reporting of Results
- Grade 9 Social Studies 1983 Achievement Test

R. Runte reviewed the test development process as used in the development of the 1983 Social Studies 9 Achievement Test.

3. SCORING PROCEDURES

D. Wasserman favored reporting separate scores for the multiple choice and written expression sections of the Social Studies 9 test.
M. Kozlow stated that since the test is not measuring writing skills per se, it is reasonable to report one score because the test measures the same objectives using different formats.

W. Unruh noted that the objective for writing an essay in social studies is not clear. He cautioned that students may be misled by the format; they may interpret the written part to mean that "writing" is the objective being evaluated.
He also suggested that a less costly, more efficient way of testing this objective could be accomplished by using a different format (e.g., fill in the blank).

R. Runte informed the committee that the Nedelsky method would be used for setting standards for the test.
J. Holmes expressed concern that the Nedelsky method could not be applied by someone not closely tied to curriculum and students, so the Technical Review Committee is not the appropriate place for setting standards.

L. Tymko asked for closure on discussion on the test development process.

The committee broke for lunch from 11:45 a.m. to 1:00 p.m.

* * * * *

L. Tymko asked for general comments re the test development process.

In general, members agreed that given current time constraints, the process is adequate. Given more lead time, however, some fine tuning should be considered, primarily in the areas of field- and pilot-testing and the structure of the TRC. The committee agreed that greater outside involvement was desirable and that some change in policy is necessary with respect to representation on the TRC.

H. Mosychuk concerned with maintaining adequate respresentation of lay people on TRCs.

W. Brooks - lack of specific purpose defined for the test.

T. Maguire: The purpose of achievement tests is clearly defined, but different interest groups tend to disagree as to what the test purpose should be. Hence, reporting is important because it should reflect and clarify the purpose of the tests. He also noted that the TRC is somewhat "light" on technical expertise.

5. STRUCTURE OF TECHNICAL REVIEW COMMITTEE

Discussion followed as to what form the TRC should take.

T. Maguire suggested that the TRC be changed to Test Review Committee with a sub-committee (Technical Advisory/Approval Committee) to review the technical aspects of the examinations. Terms of reference of the Technical Review Committee would be to check the tests for face and curricular validity.

W. Brooks proposed that two separate bodies be established with overlapping commitments; organize a pool of experts which you can draw upon as needed.

More teacher involvement, even at the risk of greater threat to test security, would eliminate the suspicion that test are biased and would thereby eliminate questions re the validity of test content.

6. ITEM DEVELOPMENT, FIELD-TESTING, PILOT-TESTING

It was recommended that the "research arm" of the SEB look at the following, time permitting:

- best sampling methods
- alternatives for the test development process (e.g., how many items do you need to test an objective and have validity)

It was recommended that the final form of the examinations be vetted and approved through the Test Review and Technical Approval Committees.

7. ADMINISTRATION, SCORING, SETTING STANDARDS, ANALYSIS, INTERPRETATION, REPORTING, USE OF RESULTS

J. Holmes stated that when addressing standards, we must understand the skills of the test-takers so that while we have provincial input, we should also look at jurisdiction results.

W. Unruh suggested that one alternative solution would be to administer say 60 provincially set multiple choice items, and time permitting, add 20 items at the local level to allow for curricular flexibility.

8. EXAM SECURITY

There was discussion re the trade-off between keeping examinations secure to allow for comparability over time, and the advantages of making them available to teachers. The committee agreed that given a choice, they would tend to opt for making the examinations public for these reasons:

(a) it would keep the tests "honest"
(b) it would benefit the systems because teachers would learn how to construct good tests, which would help to raise standards

9. TECHNICAL INPUT

W. Unruh asked what type of technical input the research arm of the SEB would like from external members on the committee.
D. Wasserman indicated that input was required in terms of:

- type of item analysis recommended
- sample sizes for field- and pilot-testing
- procedures to be used in pilot-testing
- standards and setting of standards
- competency/norm-referenced testing
- test equating
- reporting of results

(a) pass/fail basis
(b) percentage scores
(c) by section (i.e., reporting reading/writing as separate scores)

He asked the committee what they considered an ideal way to go about this and what should be accepted as minimal standards.

T. Maguire suggested that the field-testing process be a more iterative one, involving shorter tests, fewer people, and a faster turnaround time.

W. Unruh inquired as to whether any research had been done on alternative item forms. M. Kozlow replied that consideration had been given to this, but time constraints mitigated against using this format at this time.
W. Unruh suggested that the work be contracted out, if the research team of the SEB is limited by time constraints. The committee agreed that consideration should be given to the feasibility of alternate item forms.

10. T. Maguire reviewed the main points of his paper "Standards & Guidelines in the Assessment of Student Achievement" which had been circulated for discussion at the next meeting.

The committee agreed on the agenda for Meetings #3 and #4, to be held March 14 and 25, 1983 respectively.

The committee agreed that the final report to the Director should present a list of questions followed by recommendations.

The meeting adjourned at 3:30 p.m.

AD HOC STUDENT EVALUATION TECHNICAL ADVISORY COMMITTEE

NOTICE OF MEETING

Monday, February 21, 1983 at 9 a.m.

Student Evaluation Branch Boardroom
Harley Court, 4th Floor
10045 - 111 Street
Edmonton, Alberta

Phone: 427-2948

AGENDA

1. Minutes of the previous meeting

2. The Revised Student Evaluation Policy Discussion Paper, and the purposes of testing

3. The Test Development Process as used in the development of the 1983 Social Studies 9 Achievement Test.

Information Items

1. Grade 9 Program of Studies - Social Studies

2. Curriculum Specifications - Grade 9 Social Studies

3. Test Blueprint - Grade 9 Social Studies

Minutes of the Second Meeting of the AD HOC STUDENT EVALUATION TECHNICAL ADVISORY COMMITTEE

held in the Boardroom of
Harley Court, 4th Floor
10045 - 111 Street, Edmonton, Alberta
February 21, 1983

Members Present

J. Lawrence Tymko (Chairman) (SEB)
Willard M. Brooks (ATA)
Janelle Holmes (Calgary Board of Education)
Michael Kozlow (SEB)
Thomas Maguire (U. of A.)
Harry Mosychuk (Edmonton Public School Board)
Walter M. Unruh (U. of C.)
David Wasserman (SEB)

Absent

Larry Ferguson (ASTA)
Eric Mokosch (U. of L.)

Participants

Lloyd E. Symyrozum (Director, SEB)
Robert Runte (Test Development Specialist, Social Studies)

Recorder

Patricia Ford

The meeting was called to order at 9:00 a.m.

1. THE REVISED STUDENT EVALUATION POLICY DISCUSSION PAPER

L. Symyrozum joined the meeting and asked external members on the committee for feedback regarding the discussion paper.

H. Mosychuk: questioned the "purpose" section of the paper, noting that the testing program should encompass all aspects of testing and should therefore address Appeal Examinations. There is no provision for Appeal Examinations in the "purpose" section.

He requested clarification re the Provincial Role, pointing out that comprehensives essentially certify that students have acquired certain knowledge and skills as reflected in teacher-assigned marks. To reflect this, he recommended that the statement

> 1. *Certifying the basic skills and knowledge attained by students graduating from high school through the awarding of a High School Diploma.*

be changed to read

> 1. *Certifying the basic skills and knowledge attained by students graduating from high school through compulsory testing.*

J. Holmes: does not dispute the underlying premises, but feels the policy is determining ends. If ends are not met, do we go back to changing policy. Also feels there is no provision for individual needs at the local or jurisdictional levels.

H. Mosychuk: recommended that policy specify what portion of the student population is affected and what portion of that population is expected to pass and qualify for the High School Diploma. *We should be setting standards.*

W. Unruh: committee has to clarify whether this is a competency-based approach. It is important that standards be set and it is important to clarify the purpose of the marks. He emphasized the need to ensure that methods and design fit the goals of the particular test so as not to confound methodology and context.

The committee agreed that standards should be set in place in this regard.

The committee adjourned for coffee from 10:00 to 10:15 a.m.

.../2

2. TEST DEVELOPMENT PROCESS

Several documents which had been circulated were presented as information items:

- Item Control Sheet
- Test Blueprint
- Proposed Categories for Reporting of Results
- Grade 9 Social Studies 1983 Achievement Test

R. Runte reviewed the test development process as used in the development of the 1983 Social Studies 9 Achievement Test.

3. SCORING PROCEDURES

D. Wasserman favored reporting separate scores for the multiple choice and written expression sections of the Social Studies 9 test.
M. Kozlow stated that since the test is not measuring writing skills *per se*, it is reasonable to report one score because the test measures the same objectives using different formats.

W. Unruh noted that the objective for writing an essay in social studies is not clear. He cautioned that students may be misled by the format; they may interpret the written part to mean that "writing" is the objective being evaluated.
He also suggested that a less costly, more efficient way of testing this objective could be accomplished by using a different format (e.g., fill in the blank).

R. Runte informed the committee that the Nedelsky method would be used for setting standards for the test.
J. Holmes expressed concern that the Nedelsky method could not be applied by someone not closely tied to curriculum and students, so the Technical Review Committee is not the appropriate place for setting standards.

L. Tymko asked for closure on discussion on the test development process.

The committee broke for lunch from 11:45 a.m. to 1:00 p.m.

* * * * *

L. Tymko asked for general comments re the test development process.

In general, members agreed that given current time constraints, the process is adequate. Given more lead time, however, some fine tuning should be considered, primarily in the areas of field- and pilot-testing and the structure of the TRC. The committee agreed that greater outside involvement was desirable and that some change in policy is necessary with respect to representation on the TRC.

H. Mosychuk concerned with maintaining adequate representation of lay people on TRCs.

W. Brooks - lack of specific purpose defined for the test.

T. Maguire: The purpose of achievement tests is clearly defined, but different interest groups tend to disagree as to what the test purpose should be. Hence, reporting is important because it should reflect and clarify the purpose of the tests. He also noted that the TRC is somewhat "light" on technical expertise.

5. STRUCTURE OF TECHNICAL REVIEW COMMITTEE

Discussion followed as to what form the TRC should take.

T. Maguire suggested that the TRC be changed to Test Review Committee with a sub-committee (Technical Advisory Committee) to review the technical aspects of the examinations. Terms of reference of the Test Review Committee would be to check the tests for face and curricular validity.

W. Brooks proposed that two separate bodies be established with overlapping commitments; organize a pool of experts which you can draw upon as needed.

More teacher involvement, even at the risk of greater threat to test security, would eliminate the suspicion that tests are biased and would thereby eliminate questions re the validity of test content.

6. ITEM DEVELOPMENT, FIELD-TESTING, PILOT-TESTING

It was recommended that the "research arm" of the SEB look at the following, time permitting:

- best sampling methods
- alternatives for the test development process (e.g., how many items do you need to test an objective and have validity)

It was recommended that the final form of the examinations be vetted and approved through the Test Review and Technical Advisory Committees.

7. ADMINISTRATION, SCORING, SETTING STANDARDS, ANALYSIS, INTERPRETATION, REPORTING, USE OF RESULTS

J. Holmes stated that when addressing standards, we must understand the skills of the test-takers so that while we have provincial input, we should also look at jurisdiction results.

W. Unruh suggested that one alternative solution would be to administer approximately 60 provincially set multiple choice items, and time permitting, add 20 items at the local level to allow for curricular flexibility.

8. EXAM SECURITY

There was discussion re the trade-off between keeping examinations secure to allow for comparability over time, and the advantages of making them available to teachers. The committee agreed that given a choice, they would tend to opt for making the examinations public for these reasons:

(a) it would keep the tests "honest"
(b) it would benefit the systems because teachers would learn how to construct good tests, which would help to raise standards

9. TECHNICAL INPUT

W. Unruh asked what type of technical input the research arm of the SEB would like from external members on the committee.

D. Wasserman indicated that input was required in terms of:

- type of item analysis recommended
- sample sizes for field- and pilot-testing
- procedures to be used in pilot-testing
- standards and setting of standards
- competency/norm-referenced testing
- test equating
- reporting of results

  (a) pass/fail basis
  (b) percentage scores
  (c) by section (i.e., reporting reading/writing as separate scores)

He asked the committee what they considered an ideal way to go about this and what should be accepted as minimal standards.

T. Maguire suggested that the field-testing process be a more iterative one, involving shorter tests, fewer people, and a faster turnaround time.

W. Unruh inquired as to whether any research had been done on alternative item forms. M. Kozlow replied that consideration had been given to this, but time constraints mitigated against using this format at this time.

W. Unruh suggested that the work be contracted out, if the research team of the SEB is limited by time constraints. The committee agreed that consideration should be given to the feasibility of alternate item forms.

10. T. Maguire reviewed the main points of his paper "Standards & Guidelines in the Assessment of Student Achievement" which had been circulated for discussion at the next meeting.

The committee agreed on the agenda for Meetings #3 and #4, to be held March 14 and 25, 1983 respectively.

The committee agreed that the final report to the Director should present a list of questions followed by recommendations.

The meeting adjourned at 3:30 p.m.

ACHIEVEMENT TEST DEVELOPMENT PROCEDURES

Generally, a similar pattern of test development is followed by all Test Development Specialists within the Humanities Section. There are some minor variations in the pattern of development or in the number of teachers involved in the construction and revision committees. The general pattern involves the following steps:

Step 1: PREPARE CURRICULUM SPECIFICATIONS

The Curriculum Specifications for each program to be evaluated are based upon the *Program of Studies* as prescribed by the Minister of Education. The content areas and skills to be emphasized at each grade level are decided upon by a group of teachers, subject supervisors and university personnel under the direction of the Curriculum Branch of Alberta Education.

Step 2: PREPARE TEST SPECIFICATIONS AND AN INTERIM BLUEPRINT

After a decision has been made as to what tests are required, test specifications and an interim blueprint, based on the Curriculum Specifications, are developed by the Test Development Specialist. Test specifications and interim blueprints identify the key areas to be tested and the number of questions that should be attached to each objective. Generally, each reporting cell of the blueprint has at least six questions testing a particular objective in order to achieve test reliability. Ideally, there is a range in difficulty within each cell from .30 to .85, with a mean that is close to the normal intended test mean of 62.5%. (An item difficulty score represents the proportion of students who chose the correct alternative for that item. For example, a difficulty of .70 shows that .7 (70%) of the students chose the correct alternative.)

Also, a minimum of 10 items or 10% of the total test, whichever is more, are repeated from administration to administration to allow for comparability of results. The repeated items are a representative sample of difficulty levels and blueprint classifications.

TEST RELIABILITY is an important consideration in test construction. A test is said to be reliable when it measures consistently from one administration to another. Since reliability depends on item design and on the number of items, the question for the achievement tests is:

> ARE THERE ENOUGH QUESTIONS TESTING THE SAME OBJECTIVE TO ENABLE US TO CLAIM THAT THE TEST IS GIVING US AN ACCURATE ASSESSMENT OF PERFORMANCE ON THAT OBJECTIVE EACH TIME WE TEST FOR IT?

TEST VALIDITY is another important consideration in test construction. A test is valid to the extent that it measures what it is intended to measure. As the achievement tests are intended to measure the most important aspects of the curriculum, the question becomes:

> DOES THE TEST MEASURE THOSE OBJECTIVES CONSIDERED TO BE MOST IMPORTANT IN THE CURRICULUM?

This consideration is also important in classifying items into cells - that is, does each item measure the particular objective it is supposed to measure?

Another important procedure that is carried out during this preliminary planning step is the specification of item formats. Item format has implications for the type of scoring that will be required and the types of achievement that can be measured. The tests in Language Arts and Social Studies have a multiple-choice machine-scored section and a written portion that is scored by markers.

The selection of items is determined by the test objectives and is primarily the responsibility of the Test Development Unit. An attempt is made to have the blueprint match the Curriculum Specification emphases as closely as possible.

After agreement within the Student Evaluation Branch, the interim blueprint is brought before an inter-branch committee consisting of the Associate Director of Curriculum (or Associate Director of Language Services) Field Services consultants, and the members of the test development team.

Step 3: APPROVE BLUEPRINT

The blueprint is then reviewed and recommended for approval by a Technical Review Committee. This committee makes recommendations to the Director of Student Evaluation. Its role is advisory in nature.

ACHIEVEMENT TEST TECHNICAL REVIEW COMMITTEE

Membership is as follows:

1. A chairman appointed by the Director of Student Evaluation
2. A member nominated by the Associate Deputy Minister of Program Development
3. A member nominated by the Associate Deputy Minister of Program Delivery
4. A selected representative of post-secondary institutions
5. Two members nominated by the Alberta Teachers' Association who have demonstrated an interest and expertise in the subject area, but who are not currently engaged in teaching the specific course being examined
6. A member nominated by the Conference of Alberta School Superintendents (CASS)

The Technical Review Committee considers the general level of achievement expected for the test. If the test is to contain minimum competency items, then the items with an expected frequency of students achieving at the mastery level are identified.

These items are not included with the others when aiming for the normal test mean of .625.

Step 4: DEVELOP TEST ITEMS

A. Check Available Items Against Blueprint

Items from earlier tests and from other item banks may be available for use. These items are checked to determine which ones fit the test blueprint and the item format(s) being used. Items meeting these criteria can be considered for use. This selection is a Test Development Unit responsibility.

B. Identify Potential Item Writers

Experienced subject area teachers are a source of item writers. These are identified by Test Development Specialists from personal knowledge, by professional organizations, by references from regional office curriculum specialists, and by superintendents, among others. An attempt is made to have a representative sample of teachers from both urban and rural areas, from public and separate systems, and from all regions of the province. This operation is primarily a Test Development Unit responsibility.

C. Contact Superintendents Re: Item Writers

The protocol for the involvement of teachers in item writing requires that approaches be made through jurisdiction superintendents. Approval for release time is obtained through the superintendents.

D. Arrange Release Time

For item writing to be done in committees, mutually satisfactory meeting times must be arranged. These arrangements for release time are made with the superintendents, principals, and individual teachers involved. The involvement of classroom teachers through committee work helps to ensure that the items have face validity (classroom) even in the initial stages of development.

E. Hold Item Writing Sessions

Item writing sessions are held under the supervision of a Test Development Specialist. Where necessary, item builders are trained in the principles of item construction. Once they have had some in-service training, teachers work as a group or individually and meet to review and revise the items they have built. It is the responsibility of the Test Development Specialist to identify the items required to fill the cells of the blueprint and to test all objectives specified in the blueprint.

F. Screen New Items

New items built by the item building committees are screened by subject area Test Development Specialists for format, face validity, blueprint category, and other design considerations. Unsuitable items are revised where possible or replaced by new items.

During the item development process, copyright approval is sought as required for such materials as literary passages, political cartoons, and graphs and charts. If possible, art work is completed during this time as well.

Step 5: FIELD TEST THE ITEMS

Items are then field tested on a minimum of 150 students in order to determine how the items "behave" or test out in an actual testing situation. Field tests may cover select parts of the curriculum or experimental item types. The results of the tests are then analyzed and the statistics or item analyses scrutinized in order to determine whether the items require further revision. All items are classified into four categories:

A. Totally Acceptable (TA)

B. As Is (AI)

C. Revise (R)

D. Discard (D)

Between one-third and one-half of the items on a given field test are usually TA or AI.

Items requiring further changes are revised and resubmitted for further field testing.

Step 6: CONSTRUCT TEST

After items come back from field testing, the acceptable ones (TA and AI) are put together to form a test which resembles closely the final test in sample characteristics, content, organization, administration, and analysis. Normally, items range in difficulty from .30 to .85, with an overall test mean of .625.

Step. 7: APPROVE TEST

The Technical Review Committee reviews a representative sample of questions being developed for the test, commenting on their format, content, difficulty level, and discrimination index. It is at this point that the Technical Review Committee sets standards of achievement appropriate for the specific test to be administered. They also review and recommend for approval the final draft of the test, the administration instructions, and the marking standards for written response.

Step 8: ADMINISTER PILOT TEST

This may be considered the 'trial run' of the final test. To serve as an adequate guide to test reliability and validity, the pilot testing situation must duplicate as closely as possible the final testing situation. A pilot test is as close in construction to the final test as possible.

Step 9: PREPARE AND ADMINISTER FINAL TEST FORM

Final revisions are made on the basis of information collected in the administration of the pilot test. No further scrutiny by the Technical Review Committee is required at this stage unless substantial changes are made to the original pilot test. The tests are then printed commercially and distributed for administration.

Step 10: ANALYZE, EVALUATE, AND REVIEW RESULTS

The results of the test are then analyzed. The Technical Review Committee meets to review and recommend for approval the results of the test. The standards of achievement previously set may be reconsidered in light of the results at this time.

Step 11: REPORT RESULTS

The following reports are prepared:

1. The Summary Report presents an overview of the provincial results. It is disseminated to all those listed on the standard mailing list of Alberta Education.
2. The Provincial Report provides more in-depth information on the test and the results. It is designed to assist school jurisdictions in interpreting the results.
3. The Jurisdiction Report gives the results for individual jurisdiction. It is confidential to each jurisdiction.
4. The Branch Technical File consists of data that may be used by Alberta Education in future analyses.

TASK FLOW CHART AND TIME FRAME

e.g. Grade 9 Language Arts Achievement Test

STEP	TASK	COMPLETION DATE
1	Prepare curriculum specifications	Prior to Fall, 1982
2	Prepare interim blueprint	Fall, 1982
3	Approve blueprint	Winter, 1982
4	Develop test items	Winter, 1982 & Spring, 1983
5	Field test items	Spring, 1983
6	Construct pilot test	Summer, 1983
7	Approve pilot test	Fall, 1983
8	Administer pilot test	Spring, 1984
9	Prepare and administer final test	Spring, 1985
10	Analyze, evaluate, and review results	Summer, 1985
11	Report results	Fall, 1985

GRADE 9 SOCIAL STUDIES ACHIEVEMENT TEST

BLUEPRINT FOR MULTIPLE CHOICE SECTION

OBJECTIVES	TOPIC A	TOPIC B	TOPIC C	NUMBER OF ITEMS	PERCENTAGE WEIGHTING
VALUE - Understanding Values - Develop Competencies				8	12%
SUB-TOTAL	3	3	2	8	
KNOWLEDGE (CONCEPTS)					44%
- Basic Economics		1		1	
- Centrally Planned Economy					
- Historical Evolution		3		3	
- Principles of Centralization		2		2	
- Conservation			1	1	
- Control		2		2	
- Demography			1	1	
- Industrialization					
- In Britain	4			4	
- In Canada (Geog. Factors)			1	1	
- Industry					
- Primary, Secondary, Service			3	3	
- Labor/Management Relations	1			1	
- Market Economy	2			2	
- Materialism	1			1	
- Quality of Life					
- Scarcity	1			1	
- Technological Change			2	2	
- Welfare of the State		1		1	
SUB-TOTAL	9	9	8	26	
SKILL					44%
- Inquiry					
a) Identify and Focus on Issue				6	
b) Formulate Research Questions				4	
c) Gather and Organize Data				7	
d) Analyze and Evaluate Data				3	
e) Synthesize Data				3	
f) Resolve the Issue				2	
g) Apply the Decision				1	
h) Evaluate the Decision				0	
- Participation					
- Inquiry Process					
SUB-TOTAL	8	8	10	26	
TOTAL NUMBER OF ITEMS	20	20	20	60	

ITEM	CLASSIFICATION		REPORTING	STATISTICS		AI	TA	KEY	REVISED	
NUMBER	TOPIC	OBJECTIVE	CATEGORY	DIFF	PBS				YES	NO
1	A	K II A	1	760	253	✓		C		✓
2	A	K I B	1	322	336	✓		B		✓
3	A	K II C	1	288	293	✓		C	✓	
4	A	K VI B	1	824	480		✓	D		✓
5	A	S I B	6	464	266	✓		D	✓	
6	A	S I A	5	652	287	✓		C		✓
7	A	S I C	6	601	363		✓	A		✓
8	A	S I E	7	790	455	✓		D		✓
9	A	S I C	7	730	143	R		C	✓	
10	A	S I A	6	880	326	✓		B		✓
11	A	K VII	1	674	394		✓	B		✓
12	A	K IV A	3	541	290		✓	B		✓
13	A	K III A	3	871	355	✓		D	✓	
14	A	I V B 2	3	536	379		✓	C		✓
15	A	S I C	6	687	437	✓		C		✓
16	A	K II C	3	451	319	✓		A	✓	
17	A	S I A	5	584	503	✓		C		✓
18	A	V I A	4	554	406		✓	D		✓
19	A	V I B	4	567	384		✓	B		✓
20	A	V I B	4	712	414		✓	D		✓
21	B	K I A	3	777	454		✓	D		✓
22	B	V I B	4	751	489		✓	A		✓
23	B	K II A 1	1	815	222	✓		C	✓	
24	B	K II B 3	3	425	152	✓		C		✓
25	B	K II E 3	1	635	574		✓	A		✓
26	B	K III A	3	524	465		✓	D		✓
27	B	S I C	6	571	350	✓		A	✓	
28	B	S I D	7	777	420	✓		B		✓
29	B	K III A	1	627	459		✓	A		✓
30	B	K IV C 1	3	433	343		✓	C		
31	B	K VI B	3	502	274	✓		B		✓
32	B	V I	4	609	157	✓		B		✓
33	B	S I C	6	678	463	✓		D		✓
34	B	S I D	7	425	346		✓	D		✓
35	B	S I C	6	862	262	✓		C	✓	
36	B	S I B	6	694	492		✓	B		✓
37	B	S I B	6	414	240		✓	B		✓
38	B	S I A	5	763	465		✓	A		✓
39	B	V I A	4	720	492	✓		B	✓	
40	B	K V A	3	483	384		✓	C		✓

ITEM	CLASSIFICATION		REPORTING	STATISTICS		AI	TA	KEY	REVISED	
NUMBER	TOPIC	OBJECTIVE	CATEGORY	DIFF	PBS				YES	NO
41	C	K II A	2	444	357		✓	D		✓
42	C	S I A	5	698	513		✓	D		✓
43	C	V I A	4	629	442		✓	D		✓
44	C	S I B	6	724	396		✓	A		✓
45	C	S I C	6	422	420		✓	D		✓
46	C	K III C	2	478	324		✓	B		✓
47	C	K III B	2	737	411		✓	A		✓
48	C	K I C	2	810	453	✓		A		✓
49	C	K III A	2	319	216	✓		A	✓	
50	C	K IV B 1	2	737	404		✓	B		✓
51	C	S I E	7	504	420		✓	A		✓
52	C	S I F	7	664	460		✓	C		✓
53	C	S I G	7	504	368		✓	A		✓
54	C	S I A	5	517	291		✓	B		✓
55	C	V I A	4	703	371		✓	D		✓
56	C	S I F	7	746	536		✓	C		✓
57	C	S I E	7	672	490	✓		C	✓	
58	C	K IV A	2	603	537		✓	D		✓
59	C	K V	3	NEW				B		
60	C	S I D	7	NEW				A		

OBJECTIVES	TOPIC A INDUSTRIALIZATION IN THE 19th AND 20th CENTURIES	TOPIC B CENTRAL PLANNING IN THE U.S.S.R.	TOPIC C INDUSTRIALIZATION IN CANADA	NUMBER OF ITEMS	PERCENTAGE WEIGHTING
VALUE	I A 18	I A 39	I A 43,55	8	12%
	B 19,20	B 22,32			
	II A	II A	II A		
	B	B			
	C	C			
	III A	III A	III A		
	B	B	B		
	C				
KNOWLEDGE	I A	I A 21	I A	26	44%
	B 2	B	B		
	C	C	C 48		
		D	D		
			E		
			F		
	II A 1	II A 23	II A 41		
	B 16	B 24	B		
	C 3	C			
		D			
		E 25			
	III A 13	III A 26,29	III A 49		
	B		B 47		
			C 46		
			D		
	IV A 12	IV A	IV A 58		
	B 4	B 31	B 50		
		C 30	C		
	V A	V A 40	V A		
	B 14		B		
	C		C		
			D 59		
	VI A				
	B				
	C				
	VII A 11				
	B				
	C				
SKILL	I A 6,10,17	I A 38	I A 42,54	26	44%
	B 5	B 36,37	B 44		
	C 7,9,15	C 27,33,35	C 45		
	D	D 28,34	D 60		
	E 8	E	E 51,57		
	F	F	F 52,56		
	G	G	G 53		
	H	H	H		
	II A	II A	II A		
	B	B	B		
	C	C	C		
	D	D	D		
	III	III	III		
TOTAL	20	20	20	60	100

PROPOSED CATEGORIES FOR REPORTING OF RESULTS

1. Recalls basic economic historical and geographical facts related to industrialization — 8 items

2. Recalls facts about industrialization in Canada — 7 items

3. Understands key concepts and generalizations, specifically those related to economic system and technological change, e.g., — 11 items
   - Market economies
   - Technological change
   - Economics

4. Understands underlying values and key issues in industrialization — 8 items

5. Can identify and focus on the issue — 9 items

6. Can select appropriate questions and apply data-gathering and data-organizing procedures — 11 items

7. Can analyze, evaluate, synthesize data — 10 items

8. Can apply the inquiry process

   a) Multiple choice items (see #5, #6, and #7 above)

   b) Written response:

	APPROXIMATE WEIGHTINGS
- Identification and understanding of value positions	5
- Application of understandings and significant content	10
- Presentation of a position and supporting arguments related to an issue (consideration will be given to the quality of language and expression as well)	15
	30

STANDARDS & GUIDELINES IN THE ASSESSMENT OF STUDENT ACHIEVEMENT

Thomas O. Maguire

February 1983

University of Alberta
EDMONTON, Alberta

Introduction

The results of the first round of the Achievement Testing Program were released in November of 1982. As reports were being prepared it became apparent that in their raw form, the results would be of limited use to educators across the province. What was needed was a set of guidelines to provide a basis for interpretation.

Since the goal of the Achievement Testing Program is to provide feedback to jurisdictions, the guidelines had to be made relevant to that level of aggregation. In the interpretation of test scores for individual students, there are two broad referential catagories - norm referencing and criterion referencing. Norm referencing refers to the derivation of meaning from a person's position relative to other people, and criterion referencing refers to the derivation of meaning from the tasks performed in the evaluation.

When dealing with jurisdictions, the same two referential bases could be considered. However, in the Alberta context, referring jurisdiction averages to a distribution of jurisdiction values presents serious conceptual problems. Primary among these, is the wide variation in jurisdiction size that exists across the province. At most grade levels, the ratio of the size of the largest jurisdictions to the smallest size is greater than 100 to 1. Because of their size, the very large jurisdictions are likely to be located near the centre of the distribution, whereas jurisdictions falling at the extremes will tend to have fewer students. The reason for this is fairly straight forward. It is easier to change the scores of 30 people than of 3000 people. In order for a large jurisdiction (for example 3000 students) to have an average score near the top of the distribution, it must have reasonably high scores for 3000 students. Thus, it is much more likely that a jurisdiction with only 30 students will fall near the top of the distribution of jurisdiction means.

Another problem with norm referencing is the fact that it necessarily produces winners and losers. Yet it is conceivable that the state of achievement in some curricular areas may be inadequate in all jurisdictions, or it may be very satisfactory for all jurisdictions in other curricular areas. While both of these situations are plausible neither would be identified under a norm referenced system.

These two arguments are compelling, yet there are some points in favor of using some aspects of norm referencing. About 25,000 students participate in any test administration. Their results provide a reference level against which the suitability of the test itself can be assessed. In the test development process, care was taken to produce items that accurately reflect the curriculum. Items were pretested and revised. Technical committees inspected the revised items before they were placed in the final test. In spite of the care that was taken, there were some circumstances in which items were inadequate. These inadequacies became apparent after the results were analyzed. In addition some tests turned out to be less difficult than field testing had indicated, while others turned out to be more difficult. As a consequence the performance of the population had to influence the jurisdictional assessments.

According to Nitko, (1980), there are a variety of criterion referenced systems. For the Achievement Testing program, a concerted attempt was made to have items derive from a set of well-defined curriculum specifications. (This corresponds to Nitko's second category, "Criterion-referenced tests based on well-defined but unordered domains.") The curriculum specifications were supplied by the Curriculum Branch of Alberta Education, and in most cases the importance of each objective was included in the specification. From the specifications, test blueprints were prepared and items written.

The exact process of content validation used, differed from test to test, but in all cases technical review committees examined and approved the final selection of items. After the tests had been administered the same committees assisted in preparing the reports to be presented to the various audiences. In some cases implicit interpretations of the results were made, but for the most part the reports are characterized by description rather than judgment. The prevailing belief appeared to be that the meaning of results derived directly from the kinds of tasks that the children could perform, and in that respect the referential basis was criterion referencing.

The problem with the process as described was that it did not provide any standards against which provincial or jurisdictional performances could be measured. In the case of Science it was possible to compare results on some items with results of MACOSA results. This was the exception.

To aid interpretation, sets of target intervals were developed for each of the subtests. Where possible these target intervals were based on implicit standards set by the review committees. In all cases jurisdictional averages were used to set an approximate level, which was then modified to reflect the difficulties encountered in the test administration. (For example, there were a large number of absences for the English 30 tests, and the field test sample used for English 33 was found to be inadequate).

The target intervals were set by judgment. The guiding philosophy was that if the targets were too low, they would not encourage jurisdictions to improve. If they were too high, motivation for improvement would also be attenuated. Consequently, jurisdictional averages played a significant role in setting the initial target values. So while the initial intention was to use a content or criterion referenced presentation of results, the guidelines used were strongly influenced by the comparative performances of jurisdictions.

Literature on Standard Setting

Until the 1960's the question of standard setting appeared to be directly answered in one of two ways, either by allocating a proportion of people who would pass a test, or by specifying a passing score (commonly expressed as a percent). While both procedures were arbitrary, neither was haphazard. Anyone with considerable experience in test design is capable of producing questions whose difficulty levels do not vary much from one occasion to another. In addition, with large populations the average ability level remains fairly constant over time so that whichever procedure was used, the standards were probably constant from one year to the next.

Two factors appear to have influenced standards during the 60's and 70's. The first was the general trend toward local autonomy in education which meant that classroom teachers were required to set standards at the classroom level without the benefit of knowledge of the performance of the wider population of students. The second influence was the philosophical pressure from the objectives-based testing movement. Advocates, urged teachers to abandon the practice of setting standards by comparisons with other students, and adopt criterion referencing. Properly written educational objectives were to include the standard or level of performance required for a pass.

While this analysis is over simplified, it is true that the notion of standards which had previously been accepted became a focus for debate. A special issue of the Journal of Educational Measurement (Winter, 1978), was devoted entirely to standard setting. Three articles from that issue deserve special review because they focus the debate.

In his article, Glass (1978), traced the evolution of the notion of "criterion" as "standard", pointing out that criterion referencing need not have involved the specification of a standard of performance. Behavioral objectives that state standard such as "...the student must correctly solve 7 equations within thirty minutes", were described as pseudo quantification. More importantly, they were a meaningless attempt to make objective a matter that is essentially judgmental and, in most cases, an obfuscation of what is essentially a performance comparison or norm referenced standard.

Glass identified and criticized six classes of techniques for determining criterion scores.

1. Performance of others. The passing score is set as the score achieved by a known population. For example, the Medical Council of Canada sets a standard for non-Canadians who take the licensing exams as a particular percentile of the distribution of scores made by Canadian graduates. This is clearly a norm referenced standard.

2. Counting backwards from 100%. If a particular skill is intended to be basic, then the desired performance level is 100%. However, allowance is usually made for, clerical

errors, inattention, etc., and the level is moved down to 95% or 80% or some other value that appears to have no particular justification. The 80% level seems very common among criterion referenced test advocates but no rationale appears to have been enunciated.

3. Bootstrapping on other criterion scores. While this technique has seldom been used, it is often suggested that the passing score be establised to fit the criterion of an already established competent person in the area. This is virtually identical to the problem of predictive validity in classical test theory, and merely pushes the problem of definition of competence from the original test, back to a second test.

4. Judging minimal competence. Glass discusses the various judgmental procedure for establishing standards. His review of the literature is incomplete, but the critique of the notion of "minimal competence" is compelling in that it is difficult if not impossible to define, "minimal" in any unequivocal, or psychologically meaningful sense. Nonetheless, the <u>techniques</u> described hold out some promise for the purpose of the Student Evaluation Branch.

5. Decision theoretic approaches. A great deal of time has been spent in viewing the cutting score problem as a problem in decision theory. All procedures founder on the need for having an external criterion which has a meaningful competence level. Essential this strategy, bootstrapping and operations research are special cases of the treatment of

validity in classical test theory. They eventually come down to a norm referenced base.

6. Operations research. An attempt is made to find a cutting score that maximizes some valued commodity. Here the commodity might be future achievement for example. Generally the outcomes are weighted composites since any skill leads to several others. The weights are arbitrary, and since the cutting scores will be largely dependent on the criterion weights chosen, the method appears at best, to obscure the essential judgmental nature of the criterion.

Glass's main points are that most standard setting is arbitrary, or comparative. There is no self evident standard. In his mind the most attractive alternative is to observe rates of performance and see whether they go up or down. Of course what he doesn't discuss is the need to boundaries that tell the observer when a downshift should become a matter of concern and when it is tolerable.

Scriven (1978), in the same issue of the Journal of Educational Measurement, comments on the Glass paper and notes that it is unreasonable to talk about mastery/non mastery as a black and white distinction. At the border there are areas of grey, and that while the assignment rules are arbitrary in the region of the border, they are far less arbitrary as you go away from it. In addition, he criticizes Glass for implying that standard setting is an all or nothing affair. Anyone in an applied science field ought to automatically investigate the consequences of standards in order to modify them for subsequent use. In other words, we should learn from our experience. To quote:

> To put the matter bluntly, the answer to Glass does not lie in present practice (nor does it lie in surrender) but rather in elements missing from the whole picture including better procedures for calibrating and training judges, for synthesizing subtest scores and - especially - in needs assessment.

The third useful article from the issue on standards is Hambleton (1978). He distills Glass's criticisms to two statements: (1)all methods of determining cut-off scores are arbitrary but it would be safer if less arbitrary methods were used and (2)the use of examinee gain scores is preferable because they are less arbitrary.

In Glass's article, criticism of standard setting procedures was based in part on a particular piece of research conducted by Andrew and Hecht (1976). This article (to be discussed in more detail later), described an experiment in which judges were asked to set cutting scores using two separate procedures. The results showed large differences in the cutting points set. Hambleton points out that the results are not surprising since the approaches used were conceptually different, but that different groups of judges using the same approach were remarkably consistent.

In response to the first criticism, then, Hambleton says that the research offers some hope for developing methods that are consistent and useful, albeit arbitrary to some extent. With regard to the second criticism, at the level of the individual, gain scores are remarkably unreliable, and their use at that level may be harmful.

While Glass's article makes several valid points, the rebuttals by Scriven and Hambleton seem sufficiently compelling to warrant further research into standard setting. To stake out potentially profitable ground, it is useful to review the research involving various procedures. A brief description of the most common procedures will facilitate this.

The Angoff Procedure (Angoff, 1971)

Each judge examines each item and states the probability that a "minimally acceptable person" would answer each item correctly, or, the judges could think of a number of minimally acceptable people and estimate the proportion who would get the item correct. The sum of the proportions would constitute one judge's estimate of a cutting score. Estimates are pooled across judges to reach a final result.

There are various refinements to this procedure including the use of a scale for judges' estimates (5, 20, 40, 60, 75, 90, 95), and using different summary procedures to combine judgments

(mean, median, trimmed mean, etc.)

The Ebel Procedure (Ebel, 1972)

Judges perform a series of tasks in the Ebel procedure. To begin, they classify each item into a two dimensional table of difficulty and relevance. The categories of relevance are: essential, important, acceptable and questionable. The three difficulty levels are easy, medium, and hard. After assigning each item to a cell, the judges are asked to assign a percentage to each cell indicating the proportion of items in that cell that a minimally competent candidate should be able to answer. The number of items in the cell are then multiplied by the proportions, and the results summed over cells to reach a standard.

The various modifications of the Ebel method are many. The number of categories used may vary from application to application. Sometimes pilot data are used to classify the items according to difficulty. Alternatives to the relevence continuum (eg. taxonomic level) have been used.

The Nedelsky Procedure (Nedelsky 1954)

Judges examine each alternative in each item and indicate which alternatives the minimally competent student should reject as being incorrect. The minimal passing level for that item is the reciprocal of the number of alternatives left. The standard is estimated by averaging each judge's item passing level. These values are then averaged over judges to get the final standard P. Recognizing that there would be a grey area around the standard, Nedelsky suggested that the value of P be adjusted to

$$P_1 = P + ks.$$

The value s is the standard deviation of judges' standards, and k is a constant which determines how many of the borderline candidates pass or fail. Assuming that borderline candidates have scores that are normally distributed with a mean of P and a standard deviation of s, setting k to 0 would pass one half of the borderline group, setting k to -1 would pass 84% of the borderline group.

There have been few modifications of the Nedelsky procedure. In most applications the ks term is ignored (equivalent to k=0). As we shall see in the research to be reviewed, the context in which the Nedelsky procedure is used has been varied, so that the mind set which the judges bring to the task may differ from one application to another.

<u>The Jaeger Method</u> (Jaeger 1982)

Jaeger describes his procedure as a structured, multiphase Delphi procedure. Three groups of judges were used, registered voters, high school teachers and principals and high school counselors. The process was as follows:

1. Each group was given an overview of the problem.

2. Each judge completed the test, but was not given the results until recommendations had been made. (after step 8).

3. Each judge was asked to indicate the necessity of each item, by answering the following question, "Should every regular high school graduate be able to answer this item correctly? (Yes or No).

4. The number of "yes" items was computed for each judge, and the total constituted the standard for that judge.

5. A distribution of standards was produced and given to the judges and an explanation provided to assist them in comparing their own standard to the distribution. They were also given a table containing the percentage of students who had answered the item correctly in the previous administration. They were encouraged to examine the two handouts in light of the recommendations given at stage 2.

6. The judges were then asked to rate each item (Yes, No), keeping the information that they had been given.

7. A new distribution of standards was prepared from the recommendations of the judges.

8. A third rating session was held after the judges had been given the distributions from step 7, and the distribution of test <u>scores</u> obtained by students in the previous administration. They were shown how to interpret standards in

terms of the number of students who would pass if standards were set at particular levels.

9. A final distribution of recommended standards was calculated.

An evaluation of the procedure showed little evidence of convergence from one step to the next. While the average standard shifted, the judge variation remained fairly constant. Different groups of judges varied in their estimates, with the voters recommending higher passing scores than others on reading. In mathematics, teachers and voters had almost identical distributions, and both were higher than principals and counselors.

Other Procedures

Several other procedures have been proposed (and often used) for setting standards. Many of them require the collection of achievement data from specific groups of students, and others are entirely ad hoc. Given the requirements of the achievement testing and comprehensive examinination programs, they are unlikely to be directly applicable as they stand. Nevertheless, they may play a useful part in the evaluation of other methods and are listed below for that reason.

1. Borderline group method.

Teachers are asked to select a group of students who they consider to be at the borderline. The test is administered to this group and the median score is taken to be the standard.

2. Contrasting groups method.

Teachers are asked to classify students into pass and fail groups. The test is administered and the score distributions plotted. The point at which the score distribution for the fail group crosses the distribution for the pass group is taken to be the standard.

3. The review panel method.

This method was used in many of the MACOSA tests, and is used in the B.C. assessment program. The procedure used by the B.C. Science Assessment is described here. Interpretation Panels consisting of teachers actively involved at the grade level of interest were struck. None of the teachers had been involved in the earlier activities of the assessment (except that their students may have been in the assessment sample). The panels were given the table of specifications and a copy of the achievement items classified according to the learning objectives.

Each panel member began by completing the item and setting a percentage figure for "acceptable" and "desirable" levels of performance for the province as a whole based on the percentage of students who they felt should be able to correctly answer each item. After this, they were given the results for each item in terms of the proportion of students

who answered the item correctly and then asked to rate the performances by comparing the results with their previously estimated acceptable and desirable levels. Ratings were made on a five point scale from weak to strong. An attempt was made to reach a consensus rating through discussion. At a final session panels were asked to develop rating for each domain (a set of several items testing one area) and to provide interpretive comments and recommendataions in light of provincial performances. Again, a five point scale was used.

4. There are several empirical procedures using regression and decision-theory approaches. These involve the use of an external criterion and for the reasons described by Glass, seem inappropriate.

COMPARATIVE STUDIES

There are not many studies which compare the outcomes of the various methods. A brief survey of the literature turned up four. Andrew and Hecht (1976) used the Nedelsky and Ebel procedures and individual no consensus judgments on two instruments of 90 items each made up from the odd-even split of a 180 item test used to certify professional workers. The results shown below are the average standard (%) set by the groups under the various conditions.

	Nedelsky		Ebel	
GROUP:	A(n=4)	B(n=4)	A	B
Average Individual Standard	50.3	53.7	68.8	68.0
Group Consensus	46.3	51.3	68.4	67.6

There was no significant difference between average individual ratings and group consensus. There were no differences between groups within a single procedure, but the Ebel procedure produced significantly higher ratings than the Nedelsky procedure.

Skakun and Kling (1980) compared the Nedelsky with two modifications of the Ebel procedure on an examination in medicine. In the first modified Ebel procedure, difficulty levels for each item were given (greater than .80, .30 to . 80, less than .30) as were the taxonomic levels of the items (Factural, Comprehension, Problem Solving). In the second modified Ebel procedure, taxonomic level was cross classified with relevance (Essential - failure represents a serious gap, Important - candidates should know it, Acceptable - failure does not represent a serious gap). All items had been previously categorized and the judges were required to indicate the proportion of questions in each cell that a barely qualified candidate should answer correctly. When compared to the norm referenced standard already in use, the two Ebel procedures produced similar values. The Nedelsky value was much lower. In addition, the Nedelsky standards had higher variance over judges than the Ebel standards, and the reliability (ANOVA) for the Nedelsky rating was lower.

Rock, Davis, and Werts (1980) compared Angoff and Nedelsky procedures for estimating minimal competency and average competency using four groups of judges. The Angoff procedure produced higher values than the Nedelsky procedure, there was more consistency between groups and the group took less time with the Angoff procedure. In addition, the Angoff standards for average competency was closer to the observed mean than the Nedelsky.

Saunders, Ryan and Huynh (1981) compared two approaches, both of which are modifications of the Nedelsky method. They had 180 students rate the questions on a statistics examination that they had taken previously. Procedure 1 required them to place item alternatives into categories such that students at a minimal B level could eliminate them. The score for a minimal B, was calculated as the sum of the reciprocal of the number of alternatives so classified:

$$s = \Sigma(1/n_j).$$

In procedure 2, a third category called undecided was used in addition to the other two. For the procedure

$$s = \Sigma\{1/[n_j + (u_j/2)]\}$$

where u_j is the number of alternatives in item j that the judge couldn't classify.

The mean and median passing scores were almost identical under the two conditions and both within 2 points of the passing score used by the instructor. The variation for passing scores set by the judges seemed very large (s=6, no. of items=40, mean=30). The interquartile range was about 10 points.

In all studies that compare Nedelsky and other procedures, the Nedelsky procedure produces a lower standard. The reasons for this are many. In the first place, the tasks are different. The Nedelsky focuses on items and alternatives, (i.e. parts of items), and may, to some extent, model the behavior of a student as she or he attempts a difficult item. The Ebel technique focuses on groups of items in a fairly global fashion. According to Shepard (1980), the Nedelsky procedure should not be used unless elimination of wrong answers is clearly consistent with how a minimally competent person would answer the test. She goes on to say that, "In most minimum competency testing situations items have been written so that the hardest discrimination (chosing between the right answer and the next best answer) reflects the minimum competency to be assessed." It seems likely that if items are developed so that the alternatives represent varying degress of knowing then the Nedelsky procedure, or some refinement of it would be the procedure of choice. This is particularly true when we are dealing with a test that measures a single facet of achievement. For more complex domains, the literature suggests that the Nedelsky procedure produces estimates that are undesireably low, and Ebel type procedures would appear to be more useful.

RECOMMENDATIONS

The Alberta context is not a minimum competency testing context, and so methods for setting standards have to be useful not only for determining pass-fail, but for granting special honours for excellence. Most of the procedures outlined could be modified to fit the wider purpose. Not all of the procedures apply to non multiple-choice formats, and since many of the Alberta tests involve short answer formats, some procedures (Nedelsky for example), are not applicable to some tests.

Several authors (Jaeger, 1982, and Shepard, 1976, for example), recommend iterative procedures. Judgments are fundamentally human and therefore, they are subjective. The purpose of the standard setting exercise is to provide targets for performance and reinforcement for success. The reinforcement value is at least partly related to the level of the standard, so that standards must be high enough to be reinforcing, but low enough to be attainable. Finally, standards are social and political. They must seem appropriate to the various publics: (parents, employers, students, current teachers and subsequent teachers).

Since judgments must satisfy all of these purposes, and since human assessment is fallible, it would seem appropriate to evaluate procedures that allow standards to be modified by test results. With this in mind I would recommend that the Ebel and Nedelsky procedures be evaluated by incorporating them into sequential formats like Jaeger's.

1. Panels of appropriate publics be struck.

2. Each judge takes the test, retaining the copy until the end of the task.

3. Judges are divided in Ebel and Nedelsky groups (E and N). The E groups are given classifications of items according to the table of specifications and asked to indicate the proportion of items a passing student should get correct. They are told where each item belongs in the table.

   N groups are asked to do two tasks: (a)eliminate the responses that a passing student should eliminate, and (b)give a simple yes or no as to whether a passing student would get it right.

4. E judges are fed back the difficulty levels of each item in each of the cells, the cell judgments made by other members of the group, and the correct answers. They are asked to indicate the proportion of passing students who should have got items correct in these cells. N judges are fed back the difficulty levels of each item, the numbers of students chosing each alternative, the correct answer, and the judgments made by other members of the group. They are then asked to redo the two Nedelsky tasks.

5. Revised distributions of standards are calculated by the E, N(a) and N(b) procedures. Consequence in terms of numbers of

student passing under the three procedures are calculated.

6. The data from 5 are presented to all groups together with a description of the procedures used in each case and an attempt is made to gain consensus.

BIBLIOGRAPHY AND REFERENCES

*Andrew, B.J. and Hecht, J.T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

*Angoff, W. Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement, Washington, D.C., Amerson Council in Education, 1971.

Burton, N.W. Societal standards. Journal of Educational Measurement, 1978, 15, 263-271.

*Ebel, R.L. Essentials of Educational Measurement. Englewood Cliffs, N.J., Prentice Hall, 1972.

*Glass, G.V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-262.

*Hambleton, R. On the use of cut-off scores with criterion referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-290.

Hambleton, R. and Eignor, D. Competency test development, validation and standard setting. Chapter 9 in Jaegar and Tittle, 1980.

*Jaeger, R.M. An iterative structured judgment process for establishing standards competency tests: Theory and application. Educational Evaluation and Policy Analysis, 1982, 4, 461-475.

Jaegar, R.M. Measurement consequences of selected standard-setting Florida Journal of Educational Research, 1976, 18, 22-25.

Jaegar, R.M. and Tittle, C.K. Minimum Competency Achievement Testing, Berkeley: McCutchan Publishing, 1980. Ref. Hambleton & ~~Eg~~ Jaeggar

Livingstone, S.A. and Kastrinos, W. A study of the reliability of Nedelsky's method for choosing a passing score. 198~~7~~4, Research Report RR82-6, ETS, Princeton, New Jersey.

Meskauskas, J. Evaluation models for criterion-referenced testing. Review of Educational Research, 1976, 46, 133-158.

*Nitko, A. Criterion referencing schemes. New Directions for Testing and Measurement, 1980, 6, 35-72.

*Rock, D., Werts, E., and Werts, C. An empirical comparison of judgmental approaches to standard setting procedures. Research Report RR80-7, ETS, Princeton New Jersey, 1980.

*Scriver, M. How to anchor standards. Journal of Educational Measurement, 1978, 15, 273-275.

*Saunders, J.C., Ryan, J.P., and Huynh Huynh. A comparison of two approaches to setting passing scores based on the Nedelsky procedure. Applied Psychological Measurement. 1981, 5, 209-217.

*Shepard, L.A. Setting standards and living with them. Florida Journal of Educational Research, 1976, 18, 28-32.

Shepard, L.A. Technical issues in minimum competency testing. In D.C. Berliner (Ed.) Review of Research in Education, Vol 8. Itasca, Illinois: Peacock, 1980.

*Skakun, E.N. and Kling, S. Comparability of methods for setting standards. Journal of Educational Measurement, 1980, 17, 229-235.

* these articles are referenced in the paper